Aides à la navigation dans un corpus de transcriptions d'oral

Frederik Cailliau (1, 2), Claude de Loupy (3)

(1) LIPN – Institut Galilée – Université Paris-Nord, 99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
(2) Sinequa Labs – 51 rue Ledru-Rollin, 94200 Ivry-sur-Seine
cailliau @ sinequa.com
(3) Syllabs – 3 rue Castex, c/o Agoranov, 75004 Paris
loupy @ syllabs.com

Résumé

Dans cet article, nous évaluons les performances de fonctionnalités d'aide à la navigation dans un contexte de recherche dans un corpus audio. Nous montrons que les particularités de la transcription et, en particulier les erreurs, conduisent à une dégradation parfois importante des performances des outils d'analyse. Si la navigation par concepts reste dans des niveaux d'erreur acceptables, la reconnaissance des entités nommées, utilisée pour l'aide à la lecture, voit ses performances fortement baisser. Notre remise en doute de la portabilité de ces fonctions à un corpus oral est néanmoins atténuée par la nature même du corpus qui incite à considérer que toute méthode permettant de réduire le temps d'accès à l'information est pertinente, même si les outils utilisés sont imparfaits.

Abstract

In this paper we evaluate the performances of navigation facilities within the context of information retrieval performed on an audio corpus. We show that the issues about transcription, especially the errors, lead to a sometimes important deterioration of the performances of the analysing tools. While the navigation by concepts remains within an acceptable error rate, the recognition of named entities used in fast reading undergo a performance drop. Our caution to the portability of these functions to a speech corpus is attenuated by the nature of the corpus: access time to a speech corpus can be very long, and therefore all methods that reduce access time are good to take.

Mots-clés : évaluation, moteur de recherche, corpus oral
Keywords: evaluation, search engine, speech corpus

1 Introduction

Les corpus oraux font de plus en plus partie de notre quotidien, aussi bien à travers le web que dans notre environnement professionnel. Leur taille est dans une phase de très forte croissance du fait de la généralisation des podcasts. Face à la masse grandissante de données disponibles et les grands progrès constatés dans les technologies de transcription depuis les 15 dernières années, la recherche à l'intérieur de ces enregistrements s'impose. En particulier, les techniques d'aide à la navigation appliquées aux corpus écrits devraient être particulièrement utiles du fait du temps nécessaire à l'écoute d'une émission entière. Il a été montré que les performances des outils de transcriptions ont une influence relativement faible sur les performances des moteurs de recherche sur l'audio [Allen, 2002]. Mais ces performances de transcriptions ont-elles un impact important sur les aides à la navigation ?

Cette évaluation s'inscrit dans une série de travaux menés depuis les années 90 comme la BNN (Merlino *et al.*, 1997), Speechbot (Van Thong *et al.*, 2002), SCAN (Choi *et al.*, 1999). Des outils d'aide à la navigation pour l'audio ont déjà été testés (Anick & Tipirneni, 1999) mais concernent des fonctionnalités moins évoluées que ce qui est actuellement utilisé pour l'écrit.

Le présent article se place dans le cadre et à la suite du projet AudioSurf[1] dont le but était de créer une plate-forme d'indexation de l'audio et, en particulier, un moteur de recherche sur l'audio ayant les mêmes fonctionnalités qu'un moteur de recherche sur le texte. Les moteurs de recherche sur les textes écrits ont fait de grands progrès depuis quelques années en incluant des fonctionnalités d'aide à la navigation qui permettent de donner des informations complémentaires à l'utilisateur, de lui permettre de spécifier sa requête et d'interagir avec le système.

Nous présentons ici une application du moteur Intuition de Sinequa à l'indexation de corpus oraux transcrits à l'aide de l'outil du LIMSI et de Vecsys (Gauvain *et al.*, 2000) et les conséquences des particularités de tels corpus sur les fonctionnalités d'aide à la navigation. En section 2, nous décrivons le moteur de recherche Intuition, le principe des aides à la navigation ainsi que l'évaluation de leur apport. La section 3 décrit le corpus de transcriptions, ses particularités ainsi que les implications de ces particularités sur les performances de l'outil. Enfin, en section 4, nous présentons les résultats des évaluations que nous avons menées.

2 La plateforme Intuition

2.1 Présentation

Intuition est une plateforme de recherche d'information développée par Sinequa[2], constituée d'un moteur de recherche et d'interfaces de navigation. Elle repose sur des traitements

---

[1] Le projet AudioSurf a été financé dans le cadre du Réseau National des Technologies Logicielles (appel RNTL 2002). Il avait comme partenaire Sinequa[1] (leader), la société Vecsys[1], le LIMSI[1] et le partenaire valideur Radio France.

[2] Pour plus d'informations : http://www.sinequa.com/

linguistiques, statistiques et sémantiques qui augmentent la pertinence des documents trouvés et accélèrent la recherche des utilisateurs (cf. section 2.2).

La figure 1 présente l'interface du moteur de recherche telle qu'elle a été conçue pour le corpus audio. Globalement, cette interface est similaire à celle sur les textes écrits. Certains éléments ont cependant été ajoutés comme l'accès direct à l'écoute du passage, sa durée, etc.

Sur le volet de gauche, apparaissent des listes de concepts[3], d'entités (noms de lieux, d'organisations et de personnes). Ces éléments sont contextuels par rapport à la requête (Crestan & Loupy, 2004) et permettent à l'utilisateur de la préciser. Un simple clic sur une des entités permet de relancer une requête demandant des documents répondant à la requête précédente et contenant le terme choisi. L'extraction des entités, en combinaison avec un équilibrage statistique, se transforme alors en générateur de filtres à la volée qui permet de restreindre rapidement le nombre de documents de la liste des réponses.

L'extraction des entités est faite à partir de grammaires locales écrites sous forme de transducteurs, qui prennent en compte les résultats d'un étiquetage morphosyntaxique et d'une lemmatisation. Plus d'informations sur les ressources linguistiques utilisées dans ces traitements peuvent être retrouvées dans Cailliau (2006).

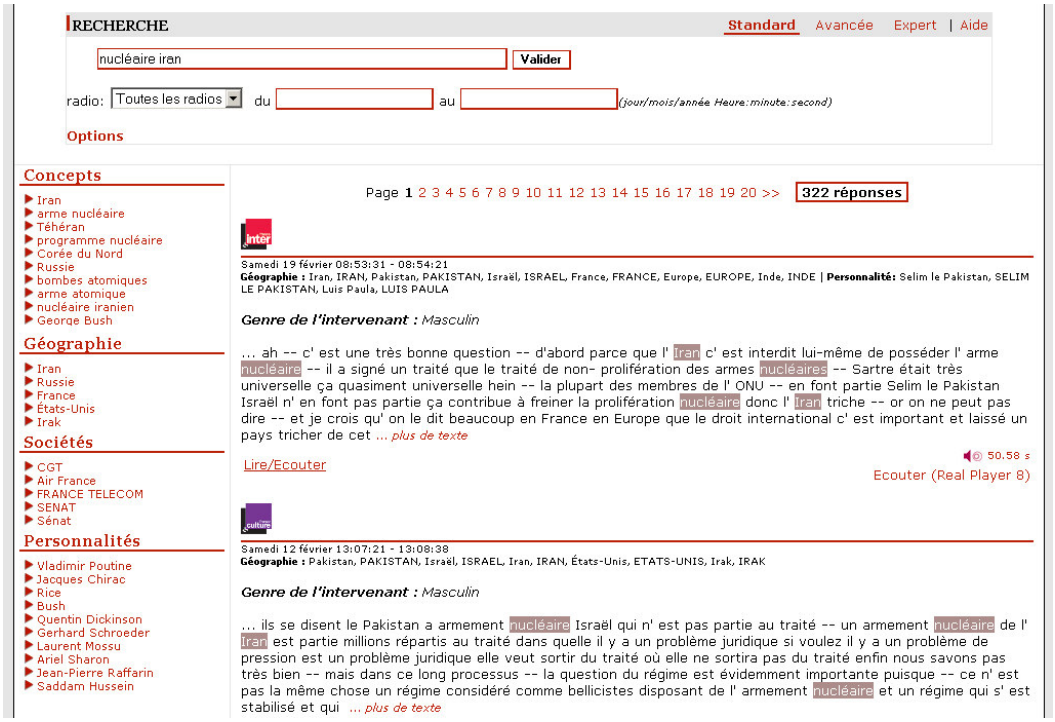

Figure 1 : Interface du moteur de recherche sur le corpus audio

L'aide à la lecture est une deuxième application des entités visible pour l'utilisateur. Elle consiste à mettre en couleur les différentes entités nommées qui ont été identifiées à l'intérieur d'un document afin de favoriser une lecture rapide par le repérage des passages importants. Par exemple, les personnes seront visualisées en rouge, les lieux en bleu, etc. Il est ainsi

---

[3] Appelés parfois aussi *termes associés*.

possible de repérer très vite ce dont parle le document. Pour l'évaluation, nous nous concentrerons sur le rappel et la précision en reconnaissance des personnes par le système.

Les interfaces décrites se complètent par un autre type de navigation, qui ne sera pas évalué dans cet article : la fonction des documents similaires. Elle permet de retrouver des documents sémantiquement proches de celui que l'utilisateur vient de regarder.

Nous évaluerons l'impact, sur ces fonctionnalités d'aide à la navigation, du passage à des corpus oraux dans la section 4.

2.2 Évaluation du principe de navigation

Le principe de navigation utilisé ici a été validé sur l'écrit [Crestan & Loupy, 2004]. Nous avons effectué une analyse mettant en jeu :

- 775 000 articles issus du journal *Le Monde* (années 1989 à 2002) ;
- 6 interfaces différentes utilisant l'une ou l'autre des fonctions de navigation ;
- 18 requêtes dont 12 de type recherche documentaire (traductions de requêtes provenant de TREC-6, ad'hoc [Voorhees & Harman, 1997]) et 6 requêtes factuelles (traductions de requêtes provenant de TREC-11, question/answering [Voorhees, 2003]) ;
- 6 personnes de formation et intérêts différents ayant pour instruction de passer exactement 10 mn par requête pour retrouver le maximum de documents pertinents. Chaque document visualisé devait être classé pertinent ou non pertinent par l'utilisateur.

Les résultats ont été très satisfaisants puisque l'interface donnant accès à toutes les aides à la navigation a permis :

- de diminuer le temps d'accès au premier document pertinent par deux en moyenne (248 s → 122 s) ;
- d'augmenter presque par deux en moyenne le nombre de documents pertinents retrouvés (3,83 → 6,56) ;
- de diminuer très significativement le nombre de documents non pertinents visualisés (7,17 → 4,28).

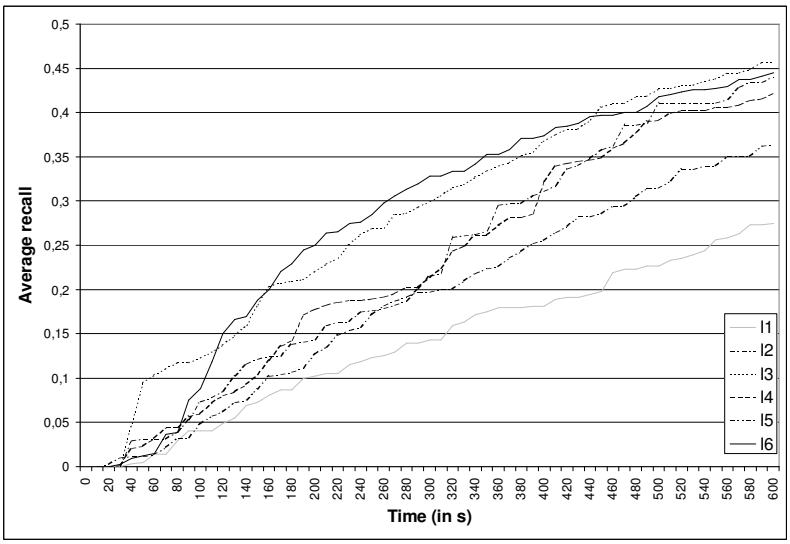

Figure 1 : Évaluation de l'apport des aides à la navigation en prenant compte du temps

La courbe précédente montre la progression du nombre de documents pertinents récupérés en utilisant les différentes interfaces. La courbe I6 (interface utilisant toutes les fonctionnalités d'aide à la navigation) obtient des résultats très au-dessus de la courbe I1 (interface basique).

Ces expériences ont donc montré l'intérêt de ces aides à la navigation sur un corpus écrit. Un corpus oral présente en revanche des difficultés pouvant réduire l'utilité de telles fonctionnalités.

3 Difficultés apportées par la transcription

Le corpus est constitué de 1048 fichiers au format xml, représentant chacun une heure de transcription automatique. Ils couvrent l'ensemble des émissions radio de France Culture et de France Inter dans la période du 6/2/05 au 28/2/05. L'unité habituelle d'indexation dans un contexte de corpus écrit est le fichier, qui correspond dans la majorité des cas aussi à une unité thématique. La notion de document a donc dû être redéfinie puisque plusieurs émissions sont présentes dans une heure de radio.

La transcription issue de l'outil de Vecsys et du LIMSI (Gauvain *et al.*, 2000) n'est pas un texte conforme à ceux habituellement traités à Sinequa. La Figure 2 montre un passage transcrit par cet outil.

Intervenante féminine

au troisième jour de sa **visite** officielle en **Chine** le président français Jacques **Chirac** a une nouvelle fois salué les relations de confiance qui unissent les deux pays -- lors d' une conférence de presse de clôture le chef de l' état a fait le bilan de son déplacement -- il a souligné la proximité de vues entre Paris et Pékin sur les grands dossiers internationaux du moment -- depuis la **Chine** le président est aussi intervenu dans le débat qui a cours actuellement dans l' hexagone concernant la retransmission télévisée des conseils des ministres une proposition formulée par le chef du gouvernement Dominique de Villepin et à laquelle le président s' est dite opposée au niveau économique et la **visite** de Jacques **Chirac** en **Chine** a permis à Airbus et à Alstom de rafler plusieurs gros contrats le chef de l' état a également posé ce vendredi -- la première pierre d' une usine PSA Peugeot Citroën avoue Anne la grande ville industrielle du centre du pays -- depuis peu Anne justement Jacques **Chirac** a évoqué le dossier du nucléaire iranien -- le chef de l' état s' est prononcé en faveur de sanctions adaptées et provisoire et irréversible contre Téhéran -- en cas d' échec du dialogue avec les Européens sur l' enrichissement de l' uranium

Intervenant masculin

il est évident -- et l' ensemble de -- la communauté internationale reconnaît notamment la Chine mais aussi la Russie -- les Européens les États-Unis -- que l' ambition clairement affichée -- par l' Iran n' est pas compatible avec une certaine idée que nous nous faisons de la non- prolifération -- en raison des dangers qu' elle représente sa propre représentera prolifération

Figure 2 : Exemple de transcription (les mots en gras sont ceux présents dans la requête d'origine)

Nous voyons ici, des caractéristiques classiques d'une transcription automatique :

– Chaque fichier xml est structuré par les tours de parole. Ceux-ci ne représentent pas forcément une unité thématique, mais ils ont été indexés comme des documents faute d'un meilleur découpage. Les balises des tours de parole comportent des attributs avec l'heure de début et de fin du tour de parole. Ces informations sont exploitées dans la maquette pour retrouver la partie du fichier audio qui y correspond. D'autres attributs non exploités sont l'identifiant de la personne qui parle et son sexe. Un flux de parole n'est pas aussi propre qu'un article de journal. Les locuteurs peuvent se couper la

parole, parler en même temps ce qui rend la transcription très aléatoire. La qualité de transcription peut être très différente d'un locuteur à un autre selon la façon d'articuler, l'accent ou le fait que le journaliste peut être sur son plateau de radio alors que l'interviewé est au téléphone (d'où une qualité de son très mauvaise).

– Le texte ne comporte aucune ponctuation. Les seules ponctuations présentes sont les deux traits qui indiquent une pause, une respiration, mais auxquels on ne peut attribuer de sémantique ou de syntaxe significative pour nos traitements. Les majuscules de début de phrase ne sont pas non plus données. L'unité phrastique est donc complètement absente et cède sa place au tour de parole. Nous avons tenté d'effectuer des adaptations de nos modèles statistiques pour prendre en compte ce phénomène, mais il aurait fallu un étiquetage de corpus pour pouvoir l'effectuer de manière correcte. Du fait de l'absence d'un tel corpus étiqueté, les expériences, utilisant des corpus normaux transformés pour ressembler à du corpus oral n'ont pas été probantes. A cause de l'absence de ponctuation et du fait de la syntaxe propre à l'oral, les transcriptions, même si elles sont de très bonne qualité, sont souvent difficilement lisibles. L'écoute des morceaux sélectionnés s'impose pour une bonne compréhension. Un tour de parole dans un journal se termine par exemple souvent par le nom de la personne qui va prendre la parole juste après dans la suite du bulletin sans aucune transition : « […] *une gauche à réunifier dans un bel ensemble Frédéric Pommier* ».

– La transcription comporte l'ensemble des disfluences, hésitations, répétitions, faux départs, etc. propres à l'oral et présente donc souvent des différences importantes par rapport à un texte écrit.

– Il y a des erreurs de transcription. La transcription de la Figure 2 est excellente mais comporte malgré tout quelques erreurs. Ainsi, à la 9ème ligne, le logiciel de transcription a écrit « *avoue Anne* » au lieu de « *à Wuhan* » (ville n'étant pas présente dans le lexique du système). Les transcriptions sont en général de très bonne qualité : le Word Error Rate (WER) sur les émissions radio a été évalué en 2001 à 20% sur le type de corpus qui nous intéresse ici (Gauvain *et al.*, 2001) mais les modèles ont été améliorés depuis et ont été évalués à 11,9% de WER pendant la campagne ESTER (Galliano et al., 2005). D'après nos observations, les erreurs relevées dans le corpus donné sont dues à la présence d'un bruit ou d'une musique de fond, à des lacunes lexicales ou au non-branchement de la détection de la langue. Pour ce dernier cas, il arrive qu'une personne parle en anglais et qu'un interprète effectue alors la traduction. Un grand nombre d'erreurs est alors généré. Dans un esprit un peu différent, les chiffres peuvent être transcrits en lettres, ce qui est le cas pour certaines années ou dans l'exemple suivant : « […] *une petite baisse de zéro zéro neuf pour-cent à quatre mille cinq points* […] ». Il est bien sûr possible de traiter facilement ce dernier point mais des particularités de ce type impliquent des ajouts de modules.

L'ensemble de ces points conduit à un certain nombre d'erreurs et de problèmes pour les fonctionnalités qui suivent, en particulier les traitements linguistiques.

4 Évaluation

Nous avons mis en place la plate-forme Intuition avec un corpus oral sans aucune adaptation des traitements décrits dans 2. Nous mesurerons leurs performance et robustesse sur un corpus oral à travers deux fonctions principales d'Intuition : la navigation par les concepts et les

entités nommées d'une part et l'extraction des entités dans les documents qui servent à l'aide à la lecture d'autre part. Ce qui est évalué ici n'est pas le WER de la transcription mais son impact sur l'aide à la navigation.

4.1 Évaluation de la navigation

4.1.1 Navigation par concepts

A partir d'un jeu de requêtes existant issues de logs d'un client de Sinequa, 40 requêtes ont été sélectionnées auxquelles au moins 50 documents dans le corpus répondent. Ces requêtes, de un à quatre mots, n'ont subies aucune modification (casse, orthographe, etc.). Elles posent des questions sur des noms de personnes (*saddam* ; *mahmoud abbas* ; …), des questions sur des noms de personnes en association avec un concept (*sistani irak* ; *sharon rice paix* ; …), des questions thématiques (*fatah* ; *armes nucléaires* ; …) ou factuelles pour obtenir une information précise (*chiites élections irak* ; *tgv Paris strasbourg* ; …). Le jeu complet est présenté en annexe.

Afin de pouvoir comparer les résultats de l'évaluation sur le corpus oral aux performances posées sur l'écrit, nous avons fait les mêmes tests sur un corpus écrit de type presse, composé de 21984 fichiers xml pour une totalité de 81,1 Mo. Le corpus oral est donc celui présenté en section 3.

Nous avons évalué les concepts qui sont extraits en fonction d'une requête. Cette évaluation est basée sur leur structure, c'est-à-dire que nous avons cherché à savoir s'ils étaient bien formés. Le but de cet article étant d'analyser l'impact des particularités de la transcription, la pertinence des concepts par rapport à la requête n'est pas évaluée.

L'évaluation elle-même porte sur les 40 premiers concepts rapportés par le système. Elle a été effectuée par 3 personnes ayant une compétence en linguistique. Pour chaque concept présenté, il était demandé de noter s'il était ou non bien formé. La figure suivante montre l'évolution des erreurs dans les concepts extraits.

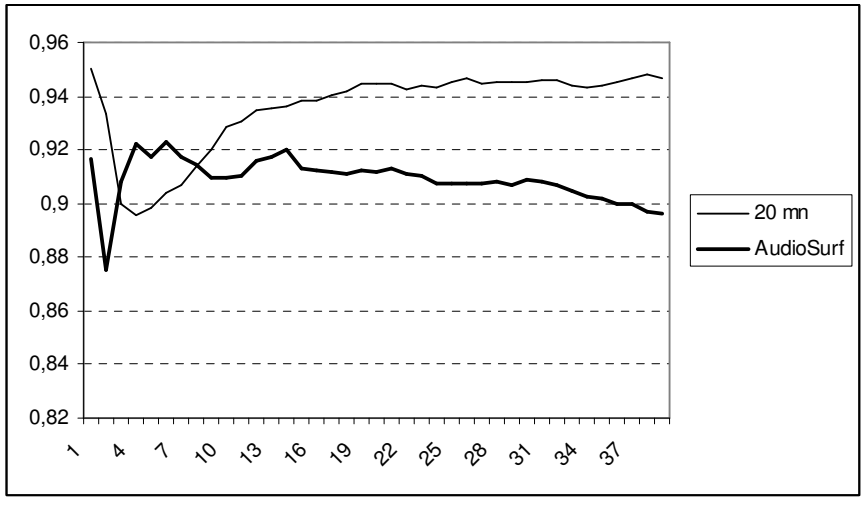

Figure 3 : Évaluation du nombre de concepts bien formés extraits d'un corpus audio (AudioSurf) et d'un corpus écrit (20 mn)

On peut voir que les concepts mal formés sont plus nombreux sur les transcriptions que sur des textes normaux comme nous pouvions nous y attendre. Le taux d'erreur moyen sur le texte est de 5% et de 10% sur les transcriptions.

Le phénomène de forte décroissance en début de courbe est dû à de mauvaises analyses des concepts et non à une mauvaise transcription (on peut voir qu'elle apparaît aussi sur le texte). Il s'agit de noms de lieux ou de personnes du Moyen Orient pour lesquels les transducteurs ne sont pas assez robustes. 16 requêtes sur 40 portent sur cette région du monde et des concepts comme *Char el* (au lieu de *Charm el Cheikh*) sont fréquemment extraits et se trouvent en tête de liste car le concept complet est très pertinent.

La croissance du nombre d'erreurs est assez forte (5 points) lorsque l'on passe aux transcriptions. Cette décroissance provient d'erreurs comme « *sa patte héros* » au lieu de *Zapatero*, « *Traque Tommy* » pour *trachéotomie* ou « *Langeais Luce* » à la place de « *l'Angélus* ». Néanmoins, nous restons dans un ordre d'erreurs acceptable (un concept sur 10 mal formé), c'est-à-dire qu'il est visible et bien présent mais non préjudiciable à la navigation.

4.1.2 Navigation par entités nommées

L'évaluation de la navigation par entités nommées n'était pas pertinente dans le contexte présent. Les lieux et les entreprises sont extraits par listes (avec quelques contextes restrictifs). Sur les entreprises, seules des choses correctes sont renvoyées et les manques viennent plutôt de l'incomplétude des listes utilisées. Pour la géographie, le rappel et la précision sont très bons mais certaines erreurs récurrentes apparaissent avec « France deux » et « France inter » dont le premier terme est reconnu comme le pays.

4.2 Évaluation de l'aide à la lecture

Cette évaluation est semblable à l'étude qu'ont faite Kubala *et al.* (1998). Les émissions de France Culture comportant peu d'entités nommées, nous avons choisi comme échantillon deux heures d'émission de France Inter. L'étude a porté sur la première demi-heure de chacune de ces émissions, car c'est la partie qui est la plus dense en entités nommées.

L'identification des entités nommées est fortement liée à la présence de ces entités dans les lexiques utilisés pour la reconnaissance de la parole. Si le nom propre est inconnu de ces lexiques, les mots en question sont remplacés par des mots communs, ce qui rend impossible toute détection par les grammaires d'extraction.

Le tableau suivant présente une évaluation du rappel et de la précision de la reconnaissance des personnes dans 3 contextes :

- La transcription automatique c'est-à-dire sans se préoccuper des mauvaises transcriptions. Ainsi, si une personne est citée à l'oral mais que la transcription en la fait pas apparaître (elle se trompe), elle n'est pas prise en compte dans le calcul.
- La confrontation à l'oral : si une entité est mal transcrite, elle sera comptabilisée quand même, ce qui fait chuter le rappel. Nous avons donc corrigé la transcription automatique pour effectuer cette évaluation.
- La transcription manuelle : afin d'évaluer l'impact des erreurs de transcription, nous avons corrigé manuellement celle-ci et repassé l'extraction des entités afin de réévaluer la précision et le rappel sur une transcription jugée sans erreur par le transcripteur humain.

	transcription automatique	confrontation à l'oral	transcription manuelle
Précision	0,90	0,90	0,91
Rappel	0,73	0,65	0,80

Les erreurs de transcription mises en cause concernent des entités comme Mahmoud Abbas qui sont reconnues sous un autre nom (« *le dirigeant palestinien Marc Mbouda basses annoncent* »), des configurations différentes d'écriture comme pour « Jean Paul deux » où ce cas d'écriture n'a pas été prévu dans les transducteurs, etc.

On constate que les erreurs de transcription ont pour conséquence une chute du rappel de 15 points par rapport à une transcription manuelle. Ce chiffre est très important et nous conduit à penser que l'utilisation de cette fonctionnalité sur un corpus oral n'est peut-être pas pertinente. Néanmoins, les transcriptions ne sont pas faites pour être lues mais plutôt pour déterminer de quoi parle un texte et si l'on veut aller plus loin en écoutant l'émission ou non. Tous les éléments permettant d'aider l'utilisateur à appréhender plus vite l'intérêt d'une émission sont intéressants dans ce contexte. Il faudrait une évaluation de navigation avec utilisateur comme celle présentée en section 2.2 pour pouvoir conclure.

5 Conclusion et perspectives

Le but de notre étude était d'évaluer la portabilité sur du corpus oral des traitements faits habituellement sur l'écrit. En ce qui concerne la navigation par concepts, nous avons constaté une dégradation significative mais tout à fait acceptable au perçu des utilisateurs. Les performances de l'extraction des entités nommées sur les documents du corpus oral sont bien faibles en rappel, mais la précision et le rappel sont en même temps déjà un apport pour la fonctionnalité visée. Dans les systèmes qui traitent de l'oral, toute amélioration qui réduit le temps d'accès à un morceau précis est un gain pour l'utilisateur. D'autres expériences qui mettent l'utilisateur au centre de l'évaluation sont à mener, justement pour mesurer si l'apport en efficacité est comparable à celui constaté sur le texte.

Remerciements

Les auteurs tiennent à remercier Mélodie Soufflard pour son travail d'analyse et d'étiquetage.

Références

ALLEN J. (2002). Perspectives on Information Retrieval and Speech. In Information Retrieval Techniques for Speech Applications, Coden, Brown, and Srinivasan (Eds.)..

ANICK, P.G., TIPIRNENI, S. (1999). The paraphrase search assistant: terminological feedback for iterative information seeking. Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval. SIGIR '99. ACM Press, New York, NY, pp. 153-159.

CAILLIAU F. (2006). Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue. Actes de TALN 2006.

Choi J., Hindle D., Pereira F., Singhal A., Whittaker S. (1999). Spoken content-based audio navigation (SCAN). Proceedings of the ICPhS-99.

GALLIANO S., GEOFFROIS E., MOSTEFA D., CHOUKRI K., BONASTRE J.-F., GRAVIER G. (2005). The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. Proceedings of the European Conf. on Speech Communication and Technology.

GAUVAIN J.L., LAMEL L., ADDA G., ADDA-DECKER M., BARRAS C., CHEN L., KERCADIO Y. DE (2001). Processing Broadcast Audio for Information Access'. ACL 39th annual meeting, pp. 2-9.

GAUVAIN J.L., LORI L., ADDA G. (2000). Transcribing broadcast news for audio and video indexing. Communications of the ACM, vol. 43, n° 2, pp. 64-70.

KUBALA F., SCHWARTZ R., STONE R., WEISCHEDEL R. (1998). Named entity extraction from speech. Proceedings of DARPA Broadcast News Transcription and Understanding. Workshop, Lansdowne, VA.

LOUPY C. DE, CRESTAN E. (2004). Browsing Help for Faster Document Retrieval. Actes de *Coling*.

MERLINO, A., MOREY, D., MAYBURY, M. (1997). Broadcast news navigation using story segmentation. *Proceedings of the Fifth ACM international Conference on Multimedia*, MULTIMEDIA '97. ACM Press, New York, NY, pp. 381-391.

VAN THONG J.M., MORENO P.J., LOGAN B., FIDLER B., MAFFEY K., MOORES, M. (2002). SpeechBot: An Experimental Speech-based Search Engine for Multimedia Content on the Web. IEEE Transactions on Multimedia, Vol 4, Nr. 1.

Annexe : liste des requêtes

1	nucléaire iran	15	Paris	29	tgv Paris strasbourg
2	chiites irak	16	Californie	30	chiites élections irak
3	russie gaz	17	Irlande	31	mur palestine
4	explosions de gaz Paris	18	Irak	32	cessez le feu intifada
5	attentat madrid	19	ONU	33	vote constitution européenne
6	mahmoud abbas	20	OTAN	34	fatah
7	aubenas	21	chirac en chine	35	forum mondial
8	Saddam	22	bush syrie	36	kyoto
9	Hussein	23	Poutine Rice	37	djihad
10	Chirac	24	assassinat hariri	38	armes nucléaires
11	Bush	25	moubarak et abdallah	39	chomage
12	jean paul deux	26	sida new york	40	grippe pape
13	Éyadéma	27	sistani irak		
14	Jean-Pierre Raffarin	28	sharon rice paix		